IS:

Big data :-

- 3Vs to describe definition of big data :
  - Volume : there has been a large increase of data volume
  - Velocity : data is being streamed at huge speed
  - Variety : - There are many different types of data
- 4Vs :
  3Vs + Veracity : Uncertainity of data
- 5Vs :
  4Vs + Value of data (Buisness value of data)

- Big data definition :-
  - data whose scale, distribution diversity and/or timeliness require the use of new technical architectures and analytics to enable insights that unlock new sources of buisness values

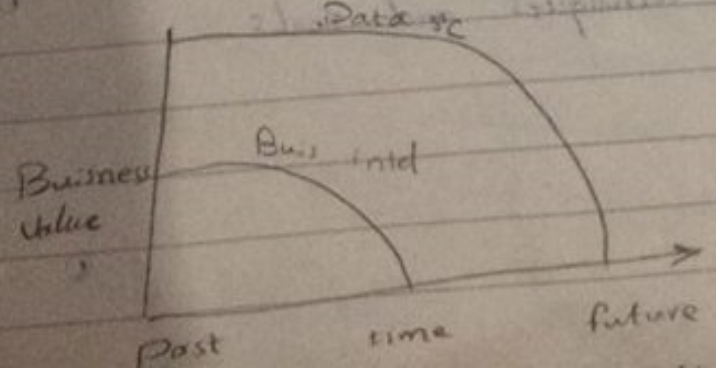
- Key characteristics of big data :-
  1- data volume : (very accelerating) Highly rate of growth
  44X increase from 2009 to 2020
  2- processing Complexity :
  - Changing data structures.
  - Uses Cases requires additional transformations and different analytical thechniques
  - Parallel Computing environments and Massively parallel processing is prefered for processing big data (enables Simultanous, parallel loading and analysis of data.
  3- Data Structure :-
  - Greater variety of data structures to mine and analyze which requires different techniques and tools

- Structured : defined data type, format, structure (OLAP)
- Semi-structured : textual data files with discernable pattern, enabling parsing (XML)
- Quasi-Structured :- Textual data with unorganized data formats Can be formatted with effort, tools and time (web clickstream data
- Unstructured : Data that has no inherent structure and usually stored as different types of files (PDFs, images and videos

* Data Repositories (analyst prespective)
  * Data Islands : - spreadsheets and Low-volume DB for record keeping
    - analyst dependent on data extracts
  * Data Warehouse: (Centralized data Containers in a purpose-built space)
    - Support BI and reporting, restricts robust analyses or exploration
    - Analyst dependent on IT and DBAs for data access
    - Analyst must spend significant time to get extracts from multiple sources
  * Analytic Sandbox: (Data assets gathered for analysis)
    - enables high performance analytics using in-db processing
    - reduces costs associated with data replication
    - Analyst-owned rather than DBA-owned
    - more robust analyses

- Buisness drivers :
People, Knowledge and conditions that intiate and support activities for which the buisness was designed

Buisness intellegance Vs Data science



- Criteria for Big data projects
  1 - Speed of decision making
  2 - throughput
  3 - Analysis flexibility

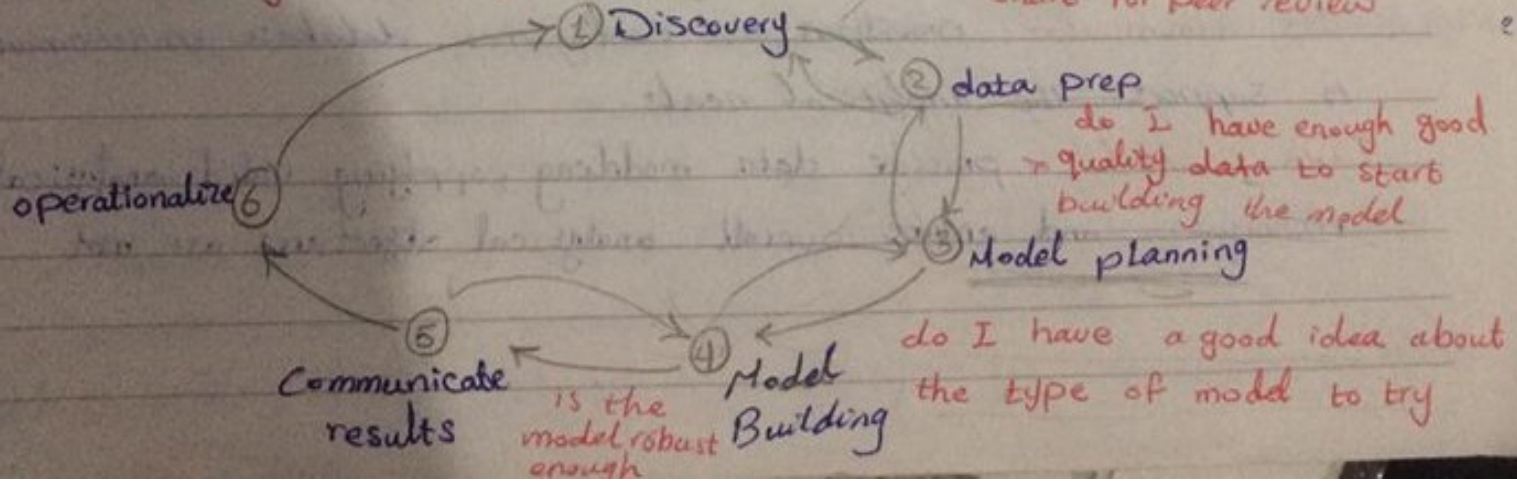
Lec 2 :-

- Three key Roles of the new data ecosystem :-
  - Deep analytical talent : (Data scientist) People with advanced training in quantitative disciplines such as mathematics and machine learning
  - Data Savvy professionals : people with basic knowledge of Statistics and/or machine learning, they can define key questions that can be answered using advanced analytics
  - Technology and data enablers: people providing technical expertise Support analytical Projects
- Data scientist key activities :-
  - reframe buisness challenges as analytics challenges
  - Design, implement and deploy statistical models and data mining techniques on big data
  - Create insights that lead to actionable recommendations
- Profile of a data scientist :-

technical - quantitative - Curious and Creative - ~~Communative~~ Communicative and Collaborative - Skeptical

Lec 3 :-

Data Analytics Lifecycle :-


do I have enough info. to draft an analytic plan and share for peer review
1 Discovery
2 data prep
do I have enough good quality data to start building the model
3 Model planning
do I have a good idea about the type of model to try
4 Model Building
is the model robust enough
5 Communicate results
6 operationalize

- Value of using the data Analytics lifecycle
  - Focus your time
  - Ensure rigidity and completeness
  - Enable better transition to members of the cross-functional analytic teams (Creating and documenting a process will help demonstrate rigor in your findings
- Need for a process to guide data science projects :-
  1- Well-defined processes can help guide any analytic project
  2- Focus of data Analytics Lifecycle is on data science projects not BI
  3- Data science projects tend to require a more consultative approach and differ from BI projects in :
     1- more due diligence in discovery phase
     2- more projects which Lack shape or structure.
     3- less predictable data
- Key Roles for a successful analytic project :-
  - Buisness user :- Someone who benefits from end results, consult and advise project team
  - project Sponser : generally provides the funding and will measure the degrees of value from final output
  - project manager :- Ensure key objectives are met on time and at expected quality
  - BI analyst :- Buisness domain expertise with deep understanding of data
  - Data engineer :- Deep technicals skills to assist with tuning SQL queries for data management
  - DB Administrator : provisions and configures database environment to support the analytical needs.
  - Data scientist :- provide data modeling, applying valid analytical techniques and ensure overall analytical objectives are met

phase 1 : Discovery :-

- Learn the buisness domain (determind amount of domain knowledge needed - determind the general analytic problem type)
- learn from the past
- Resources (People, technology, time and data)
- Frame the problem (it is the process of stating the analytics problem to be solved)
  [State the analytics problem, Why it's important; to whom - Identify key stakeholders - illustrate the current situation - objectives (goal, fail criterion) :- Identify success criteria, key risks and stakeholders (as RACI)
  RACI → way to chart responsibilities and roles in a project
  ↳ refers to people in each of 4 roles within a project
  R → responsible → actually doing the work
  A → accountable → answerable for an activity or decision
  C → Consult → typically domain experts to be consulted
  I → Inform → Individuals who need to be informed after a decision or action is taken
- Formulate intial Hypotheses
- Identify data Sources and begin learning the data

phase 2 :- Data preparation

(this step of ~~data~~ preparation is generally the most iterative and time ~~sensitive~~ intensive)

- prepare Analytic Sandbox
- perform ELT (Determine needed transformation - Determine and establish data connections for raw data)
- Execute big ELT and/or Big ETL
- Familiarize yourself with the data thoroughly (List your data sources - what's needed vs. what's available)

tools : SQL, Hadoop, mapreduce, R-package - GnuPLOT - Spotifier
data Clean visualization

- Data conditioning (clean and normalize data)
- Survey and visualize (overview, zoom & filter details-on-demand)
  → tools: R/postgresSQL, SQL, ALpine miner

Phase 3: Model planning

- Determind Methods (Select methods based on hypotheses, data structure and volume) - Ensure approches will meet buisness objective
- Techniques and workflow (Candidate tests and sequence - Identify and document modeling assumptions
- Data exploration
- Variable selection (Inputs from stakeholders and domain experts & capture essence of the predictors - Iterative testing)
- Model Sdection (Conversion to SQL or DB language for best performance - Choose techniques based on end goal)
  → tools: R, PL/R, SQL, ALpine miner

Phase 4: Model building:

- develop data sets for testing, training and production purpose
- Get the best environment you can for building models and workflows (fast hardware, parallel processing)

phase 5: Communicate the result

- Did we succeed ? did we fail
- Interpret the results
- Compare to Itl's from phase 1
- Identify key findings - Quantify buisness value

phase 6: operationalize

- run apilot - asses the benefits - provide final deliverables - Implement the model in production environment

9.

-4 Core deliverables to meet most stakeholder needs

1- Presentation for project Sponser
- "Big picture" takeaways for executive level stakeholders
- Determine key messages to aid their decision-making process
- Focus on clean, easy visuals for the presenter to explain and for the viewer to grasp

2- presentation for analysts
- Buisness process changes - Reporting changes
- Fellow data scientists will want the details and are comfortable with technical graphs

3- Code for technical people

4- technical aspecs of implementing the code

Data exploration vs. representation	
This tells you what you need to know	This tells the stakeholders what they need to know

Lec6:

Statics in the Analytical Lifecycle?
- Model building and planning
  - Can I predict the ~~outcome~~ outcome with the input that I have - which inputs?
- Model evaluation?
  - Is the model accurate?
  - Does it perform better than "the obvious guess"? another candidate model?
- Model Deployment
  - Do my predictions make a difference?
  - Are we preventing customer churn?
  - Have we raised profits

- Variance: a measure of how data points differ from the mean $s^2 = \frac{\sum (x - \bar{x})^2}{N}$ → mean

Standard deviation = Sqr root of Variance

- Hypothesis Testing: a common technique to assess the difference or significance

Null hypothesis: There is no difference

Alternate hypothesis: There is a difference.

- The aim of statical test is to determind the likehood of a value in a sample, under the assumption that Null hypothesis is true

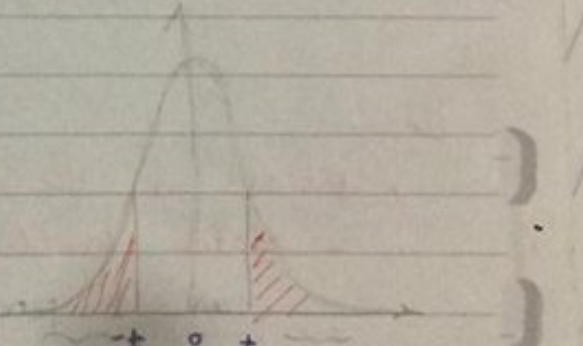
The H0 → states that there is no statistically significant difference between your sample and a reference population

The H1 → states the opposite, there is a statistically significant difference between your sample and a reference population (or between 2 samples) Slide 14

Welch's t-test:

t-statistic: $$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{S_1^2}{N_1} + \frac{S_2^2}{N_2}}}$$

p-value: area under the tails of the appropriate student distribution



if pvalue small (less than 0.05) → reject null hypothesis, assume $m_1 \neq m_2$ they are significantly different

- Significance: the probability of a false positive (α)
- Power: probability of a true positive ($1-\beta$)
- Effect Size: the size of observed difference

Hypothesis test: ANOVA

K → Population N → Samples

$$S_B^2 = \frac{1}{k-1} \sum_i n_i (m_i - m_o)^2$$

$$S_W^2 = \frac{1}{N-k} \sum_i^k \sum_j^N (x_{ij} - m_i)^2$$

$$F = S_B^2 / S_W^2 \quad \text{(test statistic)}$$

Lec 7 — Theory and methods

[I] K-means Clustering (unsupervised learning method)

finding a hidden structure within unlabeled data

Picking K: $WSS = \sum_{i=1}^{k} \sum_{j=1}^{n_i} |x_{ij} - C_i|^2$ (elbow)

$k \Rightarrow$ # of clusters n_i: # points in i^{th} cluster

$C_i \Rightarrow$ centroid of i^{th} cluster x_{ij}: jth point of i^{th} cluster

Evaluating the model:

- Do the clusters look separated in at least some of plots when you do pair-wise plots of the clusters?
  * Pair-wise plots can be used when there are not many variables
- Do you have any cluster with few data points?
  * try decreasing the value of k
- Are there splits on variables that you would expect, but don't see? * try increasing k
- Do any of the centroids seem too close to each other
  Try decreasing the value of k

- Reasons to choose (+) and cautions (-) slide 81

Lec 8

[2] Association Rules:- another unsupervised learning methods, used to discover relationships within the data (similarities)
It's not a predictive method

- The most Commonly used algorithms for association rules: Apriori
- Support - It works on frequent itemset (L) that apears together often enough
- formally often means: meets a minimum support criterion
- Support:- the precentage of transactions that contain L
- Apriori property: any subset of a frequent itemset is also frequent
- Confidence:- - Iteratively grow the frequent itemset from size 1 to size K threshold (or until we run out of support)
- Confidence :is the precentage of transactions that Contain X which also Contain Y.
- output of Apriori Algorithm : the set of all rules $X \rightarrow Y$ with min support and confidence
- Lift and leverage :-
  * Lift:- measures how many times more often X and Y occur together than expected if they were statistically independent (how x and y are really related)

$$Lift(X \rightarrow Y) = \frac{Support(X \cap Y)}{Support(X) * support(Y)}$$

  * Leverage:- measures the difference in probability of X, Y appearing together in the dataset compared to what would be expected if x and y were statistically independent (happening together by chance

$$leverage(X \rightarrow Y) = Support(X \cap Y) - [support(X) * Support(Y)]$$

- Apriori: reasons to Choose (+) and Cautions (-) slide (19)

6/11 Lec 9

~~Linear~~ regression :

Regression
- focus on relation between an outcome and its i/p variable
- provides an estimate of the outcome based on i/p values (simply predict the outcome)
- Models how changes in the i/p affect the outcome
- Regression techniques can be used as predictors, or classifier)

1. Linear Regression :
- used to estimate continous value (outcome) as a linear and additive fn of other variables
- a commonly used techniques for modeling a continous outcome
- I/p variables can be discrete or continous
- Model output : (set of estimated coefficients that indicates the relative impact of each i/p on the outcome / or linear expresson for estimating the outcome as fn of i/p variables

$$y' = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{p-1} x_{p-1} + \varepsilon$$

(ε : random error variables)

- linear regression is doing well if all the variables are uncorrelated
- to determind the best fitting line → choose the line that minimize the sum of the squares of the differences between the observed outcomes and the estimated outcomes this method is known as OLS (ordinary least square)
- Interpreting the estimated coefficients, b_j Slide 11, 12
- Diagnostics - Examing residual
  - residual :- * Difference between observed and estimated outcome
    * the observed values of the error term ε expressed as

$$e_i = y_i - \bar{y}_i \quad \text{for } i = 1, \dots, n$$

- errors are assumed to be normally distributed with a mean of zero. - Constant variance σ^2

Slide from (14) → (29)

Lec 10

(4) Logistic Regression :- (Binary classifier)

- used to estimate the probability that an event will occur as a fn of other variables
- Can be Considered a classifier as well (assign the class label with the highest probability
- input can be continous or discrete
- output : linear expression for predicting the log-odds ratio of outcomes as fn of drivers $\frac{P(X=true)}{P(X=false)} = \frac{P(X=true)}{1-P(X=true)}$

use cases - prefered method for many binary classification problems

- Binary classification example :- (we need a threshold here)
  The probability that a customer will churn
- Multiclass example :-
  the probability that politician will: vote yes/no/not show up to vote
- Diagnostics : Slide 10 → # 14
- Logistic Regression : reasons to Choose (+), Caustions (-)

Lec 11:

(5) Naive Bayesian classifier :- probabilistic classifier based on : bayes' Law and naive conditional independence assumptions

- determind most probable class label

Input :- Variables are discrete

Output :- - probability score : proportional to true probability
- class label : based on highest probability score

- use cases : prefered method for many text classification problems
  ex: (spam filtering , fraud detection)

Bayes' Law : $P(C|A) = \frac{P(A \cap C)}{P(A)} = \frac{P(A|C)P(C)}{P(A)}$

C → the class label A → observed object

[6] - Naive Bayesian Implementation Considerations

- Numerical underflow
  - * result from multiplying several probabilities near zero
  - * preventable by computing the logarithm of the products
- Zero probabilities due to unobserved attributes / classifier pairs
  - * result from rare events
  - * Handeled by smoothing
- Assign the classifier label, C_i, that maximize the value of

$$\left(\sum_{j=1}^{m} \text{Log } \hat{P}(a_j / C_i) + \text{Log } P(C_i) \right) \quad i = 1, \ldots, n$$

$\hat{P}$ → adjusted probab

- Diagnostics :- slide 16, 17, 18

Lec 12 :

[6] Decision tree classifier

- used for classification :
  - * returns probability scores of class membership, (well standardized like logistic regression), assign label based on highest scoring class - Some Algorithms return the most likely class
  - * regression tree : a variation for regression (return average value at every node - prediction can be discontinous at the decision boundries)

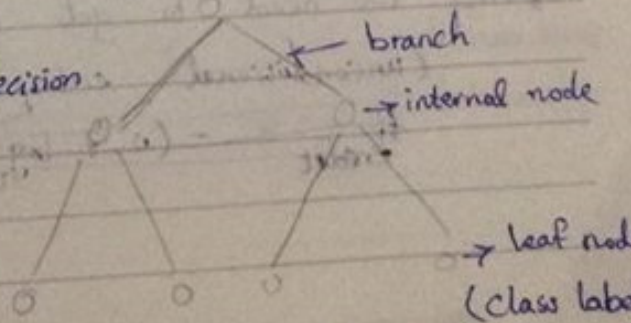
Input :- variable (Cont. or discrete)

Output :- a tree can be that decribes decision flow

- Leaf nodes return either probability score or a classification
- Trees can be converted to a set of "decision rules"

Example :- slide 5, 7

- branch : refers to the out come of decision
- Internal nodes : decision or test points each refers to a single variable



- Use Cases:
  - When a series of questions (Yes/No) are answered to arrive at a classification
  - When if-then conditions are preferred to linear models
  - Short Decision trees (where we have limited number of splits) are often used as components (weak/base learners) in ensemble techniques such as random forests
  - ensemble techniques: set of predictive models which will all vote and we take decisions based on the combination of the votes
- General Algorithm:

  to construct Tree T from training set S
  - If all examples in belong to some class in C or S is sufficiently "pure" then make a leaf C
  - otherwise: - select most informative attribute A
    - partition S according to A's value
    - recursively construct sub-trees T_1, T_2 — for subsets of S
- CART, ID3 and C4.5 are some popular Algorithm

Step 1: pick the most informative attribute

entropy-based methods are one common way

$$H = -\sum_{C} P(C) \log_2 P(C)$$

- $H = 0$ if $P(C) = 0/1$ for any class
  for binary classification, $H = 0$ is a pure node
- H is max when all classes are equally probable
  for binary classification $H = 1$ when classes 50/50

* first we need to get the base entropy of the data (unconditional entropy):

$H_{credit} = -(0.7 \log_2(0.7) + 0.3 \log_2(0.3)) = 0.88$

8.

- Conditional entropy

$$H_{attr} = -\sum_{v} P(v) \sum_{c} P(c|v) \log_2 P(c|v)$$

The weighted sum of class entropies for each value of the attribute

* Conditional entropy should be lower than Unconditional entropy

- Information Gain : the info you gain by knowing the value of an attribute $\quad$ Info Gain$_{attr}$ = H - H_{attr}

* the most Informative attribut is the attribute with highest IG

Step 2: Find partitioning with the highest Info Gain

Step 3: at each resulting node repeat step 1 and 2

(pick most ~~iterative~~ Informative attribute until node is pure enough - pure nodes → no Info Gain by splitting on other attributes)

- Diagnostics slide 16
- Reason to choose (+) & Cautions (-) slide 17
- which Classifier should I try? slide 18

lec 15

[7] Time Series Analysis

- Time Series: ordered sequence of numerical values that measured in equally spaced values over time
- Time Series Modeling: deals with time based data; involves working on time based data to derive hidden insights to make informed decision making
- Time Series Analysis: (prediction and forecasting method
It's the Analysis of Sequential data across equally spaced units of time
- Univariate time series: only one variable is measured over time
- multivariate : multiple variables are measured simultaneously